TAB A

CIA AUTOMATIC DATA PROCESSING STAFF

PROJECT [redacted]

25X1A2g

DOCUMENT/INFORMATION RETRIEVAL SYSTEM DEVELOPMENT TASK

PHASE I OUTLINE REPORT

28 June 1963

GROUP I
Excluded from automatic downgrading and declassification

CIA AUTOMATIC DATA PROCESSING STAFF

Preface

This outline report deals with the document/information retrieval system development element of Project [redacted] thinking 25X1A2g at the end of Phase I of the system development task.

The report covers:

(1) The results of [redacted] fact-finding throughout the DD/I; 25X1A2g

(2) The conclusion that a major central reference system is required;

(3) The initial concept of a new central system;

(4) A suggestion to management that a base document indexing system be urged upon the intelligence community and that this indexing function be performed once and centrally for the members of the community;

25X1A2g (5) The [redacted] plan for proceeding with the detailed development of a new document/information retrieval system (through Phases II & III);

(6) A set of general observations of particular interest to management;

(7) Major alternatives open to management; and

(8) ADPS recommendation.

25X1A2g Note: [redacted] has produced several "depth" papers for its own purposes which elaborate on the contents of this outline report. These papers are available in ADPS to persons wishing to peruse them.

CIA AUTOMATIC DATA PROCESSING STAFF

PROJECT [redacted] 25X1A2g

DOCUMENT/INFORMATION RETRIEVAL SYSTEM DEVELOPMENT TASK

* * * * *

Contents

		Page
I.	Document/Information Retrieval System Development Task	
	A. Four Phases of System Development Task.	1
	B. Phase I	
	1. Fact-Finding.	2 - 5
	2. Central vs De-Centralized System.	6 - 7
25X1A2g	3. [redacted] System Concept.	8 - 11
	4. An Intelligence Community Task, Ideally	12 - 13
	5. General Plan for Proceeding with CHIVE System Task.	14
	C. Phase II.	15 - 16
	D. Phase III	17
	E. Phase IV.	17
II.	General Observations.	18 - 22
III.	Alternatives.	23 - 24
IV.	Recommendation.	25

CIA AUTOMATIC DATA PROCESSING STAFF

PROJECT [redacted] 25X1A2g

DOCUMENT/INFORMATION RETRIEVAL SYSTEM DEVELOPMENT TASK

PHASE I OUTLINE REPORT

I. Document/Information Retrieval System Development Task

A. Four Phases of System Development Task:

Phase I - Fact-Finding and Formulation of the Overall Concept of the New System (Sept 62 - June 63)

Phase II - Detailed Systems Design (July 63 - June 64)

Phase III - Implementation of Initial Segment (July 64 - April 65)

Phase IV - Implementation of Additional Increments (May 65 - ?)

B. Phase I

1. Fact Finding

a. General

Personnel Conducting the Survey:

25X1A5a1 4 ADPS
4 [redacted] [redacted] 25X1A2g

Scope

All Offices of the DD/I
150 + components studied
Fact-finding reports prepared on each

25X1A2g Major Targets of [redacted] Fact-Finding

(1) Missions and functions of DD/I components

(2) Information sources used

(3) Internal processing and files (internal to Branch, etc. visited)

(4) Use and evaluation of external files

(5) Reports produced

(6) Information needs and problems

Survey Completed April 1963

b. Major Factors Bearing on System Development Task

Volume of Document Receipts

Multiplicity of DD/I Missions and Interests

Variety and Depth of Info Required from these Documents

Variable Time Requirements:

For basic intelligence research

For programmed, shorter-length research

For current intelligence

Trend toward Current Reporting

c. DD/I Information Resources (Present System) Composed of:

Analyst Files (para. d immediately below)

Central Info System (OCR) (para. e)

Dissemination Services (para. f)

Other Internal and External Services (para. g)

d. Analyst Files

The Analyst Files are, in fact, the primary DD/I info retrieval system in terms of:

Use rate

Response time

Indexing and content to meet analyst specifications

Uses

To check validity of new data and to determine its effect on what is already known.

To handle immediate, short lead-time ad hoc queries. Basis for more leisurely research, also.

Major Strengths

Readily accessible

Contain filtered data (reflects specialist/user judgment)

Tailored to analysts' needs (topic, sequence, and index control)

Ability to control subjects (concepts) according to the specific requirements of the analyst

Major Weaknesses

Data control largely limited to current interests

Not readily manipulated

Limited and partial historical depth

Not ideally accessible to other analysts

Organizations, personalities, areas not easily controlled

Duplicative processing among DD/I components

File maintenance detracts from analytic time

e. Central System (OCR)

General Role - Back-up to Analyst Files for:

Historical depth

Gaps in analyst file coverage

Routine, long lead time requests

Major Uses

To provide comprehensive recovery for long lead time, research projects

To provide retrieval of data not controlled in analyst files

To provide comprehensive storage and retrieval on organizations, personalities, areas

Major Strengths

Provides historical depth (institutional memory)

Comprehensive topic and area coverage

Multi-access to documents, e.g., date, source, topic, area, etc.

Backstops intelligence gaps in analyst files

Document repository

Major Weaknesses

No single point for all-source retrieval

Outputs from multiple points not compatible

STATSPEC

Insufficient emphasis given to open literature, [redacted] and cables

Slow response

Not sensitive to shifts in intelligence sources and priority interests

Inadequate geographic coordinate retrieval

Duplicative processing

f. Dissemination Services

Manual system

Minimum of 120 man years/year (rough estimate)

One million unique documents/year

10-15 million multiple copies/year

150-200 components served with specific reading requirements

General analyst satisfaction

Timely and accurate

Inefficient and costly

g. Other Information Retrieval Services

25X1A5a1

[redacted] Agriculture, etc.

Published bibliographies and indexes: Monthly Index of Russian Accessions, Referativnyy Zhurnal, ASTIA Technical Abstract Bulletin, etc.

Files of other agencies: FTD/AFSC (White Stork), Dept. of Commerce, NSA, etc.

25X1B4d
FOIAb3b1

Map Library, [redacted], NPIC, [redacted], RPB/ [redacted] RID/DDP, etc.

FOIAb3b1

Analyst chatter

2. Central vs De-Centralized System

/This is a major decision area for both systems design and management.

/A decision for a de-centralized system would mean the up-grading and coordination of the Analyst File complex with near-total dependence upon same and the correlative curtailment of the central system to a very low use, very slow response, essentially archival role.

/On the other hand, a decision for the continuation of an up-graded central system, in addition to the Analyst File system, means that heavy expenditures for a central system will not only continue but undoubtedly increase, that the effort to devise an improved central system must continue, and that eventually the resultant advanced system must be implemented and the cost and commotion of doing so accepted./

a. De-Centralized System (Analyst Files)

Pros

Provides primary support to intelligence production

Proven in practice

Reflects user needs and judgments

For majority of uses, is preferred by analysts. (Will always exist to some degree.)

Integrated sources (within clearance-level of analyst)

Cons

"Personalized" files

Difficult for others to use

Lack continuity and consistency

Difficult to manipulate

Coverage of all orgs., persons, and areas, etc. not feasible

Number and size would increase without central system

b. Centralized System

25X1A2g [redacted] concludes a central system is long-run "must" for systematic doc/info control

If improved, would:

Have higher use rate... thereby increasing the return on expenditures; and

Make inroads into present Analyst Files... thereby helping to offset costs

If accepted as a base index system for the Intelligence Community (see para. IB4 below), the [redacted] system would undoubtedly pay for itself several times over. 25X1A2g

25X1A2g 3. [redacted] System Concept

a. Very simple to say:

Central, integrated, machine-supported system to provice document and information retrieval for the total DD/I document flow.

b. Characteristics

What

All source

All geographic areas

All topics (persons, places, things, organizations, subjects)

Depth indexing

Direct entry to files (input or querying)

Single-processing of input

Single-point retrieval

Integrated output

Rapid, machine-assisted input (incl. auto indexing)

Manual indexing

Manual dissemination

Limited random access capability

Initial machine translation/Stenowriter capability

Experimental remote inquiry or display

Intermediate (1966-1967)

Large hardware complex/some advanced hardware

Manual indexing of hard copy

Some automatic indexing of machine language sources

Some character recognition (experimental)

Limited remote interrogation and display

Some automatic dissemination

Volume machine translation

Target System (1968 - ?)

Very large and advanced hardware complex, including extensive random access capability

Automatic indexing for major portions of base recovery system (incl. character recognition)

Human indexing for special info retrieval projects

Remote interrogation and display

Automatic dissemination

Volume machine translation (improved quality)

c. Elements

(1) Document storage and retrieval

(a) Persons, organizations/installations, and geographic locations to be stressed

Items of most universal interest to analysts

Weakest links in Analyst Files

Strongest elements of present Central System

Volume beyond proper handling via Analyst Files

(b) Commodities and Subjects to be covered with less emphasis

Not priority need

Limited use in central system

Analyst Files handle concepts (Subjects) better

(2) Information Storage, Manipulation, and Retrieval

(a) Correlative to Document Index System via:

Index display

Synthesis and summarization of index entries

(b) Special Projects (Language Processing), such as:

Strategic Facilities Project

25X1A2g [redacted] Project

25X1B4d [redacted]

(c) Major Automated Information System

Subject: Targets

Scope : World-wide

Inputs : Machine language files external to [redacted] 25X1A2g

: [redacted] index data (selected) 25X1A2g

: Special inputs designed for this system (For elaboration, see [redacted] X1A2g paper, same subject, dated 2 May 63)

(d) Computation (Numerical Processing), such as:

25X1B4b [redacted]

(3) Non-Literal Data Processing, such as:

25X1B5e [redacted]

(4) Machine Translation/Stenowriter

(5) Publication Support

(Use of computer for composing, type setting, etc.)

25X1A2g d. [redacted] troubled by size of task

(1) Complexity of system design

Balanced handling of such variety and volume

Accomplish objectives without undesirable consequences

(2) Hardware/software limitations

(3) Costs - personnel and budgetary

e. Full solution will require:

(1) Development of new techniques

Index, dissemination, abstract, display, input/output, etc.

(2) Development of new hardware

Memory, input/output, character readers, etc.

(3) Money and people

Major investments during developmental years.

Savings in long run?

4. Ideally, an Intelligence Community Task

a. Ideal approach seems clear - task should be done centrally for the Intelligence Community

(1) Community Effort

(a) Design and develop centrally a base doc/info system for use by community members

(b) Index centrally all docs collected/originated by Intelligence Community

Some decentralized input but conforming to base system

Some special-purpose, limited-interest categories excepted

(c) Provide base retrieval index, or suitable portions, to community members

(d) Output servicing to be performed by individual members for its local users

Base system - provided by central organization

Special files, as required - built and serviced by individual members

Some output servicing provided by central organization

(e) Initially: doc/info indexing and retrieval

(f) Eventually: translation, requirements control, etc.

(2) Executive Agent - CIA (or Intelligence Processing Center under USIB)

CIA has most suitable charter

CIA has most experience in large-scale, document systems

CIA has best/largest personnel base

CIA already started towards such a system via

25X1A2g [redacted]

[redacted]

CIA must try to do anyway for its own needs

Opportunity for CIA management to take initiative on task of vital interest to CIA, State, and DIA

Bureau of Budget should respond with real enthusiasm to such an idea (designing and developing one system instead of multiple systems; indexing centrally the intelligence document flow once instead of separately for each user). BOB could make fully ample funds available for this task and still save the taxpayer major sums of money.

b. Any such central effort for the community will take time, however.

c. CIA has need to continue [redacted] in the interim: 25X1A2g

(1) Nature of [redacted] is consistent with community idea 25X1A2g

(2) [redacted] has need for resources of community magnitude 25X1A2g

[redacted]

5. General Plan for [redacted] System Task... 25X1A2g
(July 1963 - April 1965 and continuing)

a. Design a base doc/info system for total DD/I document flow...(July 1963 - June 64 and continuing)

b. Within above context, concurrently design for implementation of the initial segment of the total system...(July 63 - June 64)

Incrementation by source

Will expand to eventual system

Keeps design tuned to real world

c. Fund and shape external R&D of hardware and software if commercial development of same is not adequate...(1963 - ?)

Must have new capabilities to accommodate growth of system

Requirements will be clarified during system design

d. Implement initial segment of new system...(July 64 - April 65)

e. Expand coverage of new system...(May 65 - ?)

C. Phase II - Detailed System Design (July 63 - June 64)

1. Personnel:

a. ADPS - continuing from Phase I

25X1A2g

b. [redacted] Contractor (114) - continuing from Phase I

c. OCR -

OLIVE has requested middle-level team from OCR to work full time on Phase II. This team would:

Receive training in EDP

Work directly with [redacted] personnel on Phase II 25X1A2g

Learn details of [redacted] system design 25X1A2g

Provide working-level OCR guidance to [redacted] 25X1A2g

Collect OCR facts/statistics required by Phase II

Serve as [redacted] liaison channels to OCR Division 25X1A2g

Become key OCR people for future operational implementation of system segments

2. Sub-Tasks of Phase II:

a. Information Processing - (CIA Team)

Coverage/scope

Index techniques

Record formats

Data reduction requirements

Query logic

Output requirements

File organizations

Etc.

b. Program Design - [redacted] 25X1A5a1

Total program concept

[redacted]

File maintenance and query programs

Utility program system

Equipment specifications

c. Hardware Study and Recommendations - [redacted] 25X1A5a1

Input/output

Micro-image store

Memory

File conversion

Machine-readable

Non-machine-readable

Reserve files

d. Design Data Collection - (CIA Team)

Data flow

Processing rates

User attitudes/needs

Present system statistics

e. Training - (CIA Team)

[redacted] staff (incl. personnel detailed from OCR) 25X1A2g

User personnel

System orientation

Task training

f. Implementation Planning - (CIA Team)

g. Management/Coordination - (JOINT Team)

D. Phase III - Initial Implementation (July 64 - April 65)

1. Sub-Tasks

a. Write programs

b. Install some additional hardware

c. Commence operation of system

Input processing

File maintenance

Output services

d. Training

2. Implement by Source Increments

Initial sources: SI and T/KH

Why SI and T/KH?

Now handled by single organizational component of OCR (Special Register). Thus:

Easier to study

Organizational dislocation within OCR resulting from implementation is minimized

Present SR system most similar to [redacted] concept 25X1A2g

Present SR responsibilities for document reference service approximate microcosm of OCR

SR personnel most familiar with machine procedures

Both sources of significant intelligence worth

E. Phase IV - Expansion of Initial Increment (May 65 - ?)

1. Addition of new sources (e.g., CS reports, OO-B's, S & T literature)

2. System configuration will change with experience

II. General Observations

25X1A2g A. [redacted] Concept

Prematurely defined at this point in time

Tentative - will change

Proportions of CAIV objectives are beyond any present-day system and beyond present-day hardware

Success is far from guaranteed

Even if many of the advanced elements of the [redacted] concept did not materialize, advantages will accrue 25X1A2g

25X1A2g B. Why a [redacted] System?

Control of more material

Deeper, more flexible index

Rapid document retrieval

Extensive information retrieval

Single service point for document retrieval

Single input processing

Integrated output

Postures the central system to grow with EDP (where future machine-support capabilities lie)

Eventual automation of some functions now done manually

C. Functions of OCR Affected/Not Affected by [redacted] System 25X1A2g

1. Affected:

Indexing and retrieval

Machine support

Dissemination

Document storage and retrieval

Photo storage and retrieval

Extracting/abstracting services

Publications procurement accounting and control

2. Not Affected:

Book cataloging and shelving

Publications and photo procurement

Library reference and circulation services (non-document)

Distribution services, i.e., the mailroom functions

Motion picture presentations

Liaison Staff

Historical Intelligence Collection

D. Organizational Effects on OCR

Interim - New system will slowly absorb people and functions

- Will constitute new element; traditional elements continue

Eventual - Present OCR Divisions will largely disappear

- Input Divisions within [redacted] will be [redacted]ized by Geographic Region 25X1A2g

- Service Division

- Systems Development Division

- Programming Division

- Computer Operations Division

25X1A2g - New non[redacted] Division(s) for non[redacted]functions 25X1A2g

E. Schedule of Effects on OCR

Phase I - Fact-Finding and Systems Concept...(Sept 62 - June 63)

Effect: None

Phase II - Detailed Systems Design...(July 63 - June 64)

Effect: None, except OCR System Trainees join with [redacted] 25X1A2g

Phase III - Initial Implementation...(July 64 - April 65)

Effect: IR index, reference, and punch personnel phase over to new system

: Conversion of old to new files accomplished (limited)

: Unconverted portions of old system serviced by EAM

Phase IV - Expansion of System...(May 65 - ?)

Effect: File maintenance/index/reference personnel from IR/BR/GR/DD/Ly (Intellofax) phase into new system

: Punch personnel in MD phase over

: File conversion accomplished (limited)

: Unconverted portions of old system to continue operations

F. Single Service Point Idea

Implementation of initial segment of [redacted] adds one more-- unless OCR develops now a single service point to tap for the consumer all pertinent OCR resources. 25X1A2g

G. Organization of OCR by Geographic Region Prior to Implementation of [redacted] 25X1A2g

Organization of OCR by Region before [redacted] implementation would foster development of single OCR service point, would lead to [redacted] increments by Region as well as source, and would facilitate successive expansions of [redacted] 25X1A2g 25X1A2g 25X1A2g

H. State-of-the-Art Implications

Conventional human indexing pushed to limit

EAM support pushed to limit

EDP offers hope through new capabilities

Even with EDP, R&D in hardware and software a "must" to expand capabilities to meet expanded [redacted] requirements in Phase IV. 25X1A2g

[redacted]

Machine indexing inferior to human indexing today

But, offers speed, consistency, and eventually perhaps comparable quality

Total document retrieval system for DD/I appears not feasible with today's equipment

Eventual DD/I system will be based on next 3-5 years of [redacted] 25X1A2g implementation experience and on R&D in industry

I. Budgetary Implications

Development and implementation costs will be heavy

Hardware Development (Government R&D support may be required)

Systems/Techniques Development (Government support almost certainly required)

Parallel Systems Operation

Conversion

Eventual system more economical per item of data controlled

J. Manpower Implications

By single input handling of documents, hope to gain manpower to permit:

Deeper indexing

Broader coverage

Greater effort on output

K. Conversion Implications

It is desirable to convert present OCR machine files, if feasible. EAM data may not be compatible with EDP files, however

--A study question for Phase II

L. Security Implications

"All-Source" clearance for all personnel operating the CHIVE system

All-Source data file:

No physical compartmentation

25X1A2g [redacted] system will have security classification code, however)